Cray Research Software Division Report

Mike Booth, Vice President, Engineering, Software and Applications Development, Cray Research, A Silicon Graphics Company, 655E Lone Oak Drive, Eagan, Minnesota 55121

ABSTRACT: *The SGI-Cray merger has enabled Cray to continue its aggressive improvement of performance and price/performance for high-end supercomputing, while maintaining source code compatibility with existing Cray supercomputer applications. This paper will cover the status and plans for Cray software on its existing and future platforms, and the plans to manage the transition to the Scalable Nodes architecture.*

1 Introduction

Cray will maintain its supercomputing leadership by continuing its focus on high-end supercomputing with aggressive improvements to performance and price/performance. As Cray makes these improvements, it will protect its customers' investments in applications, hardware, software, peripherals, and infrastructure.

The SGI-Cray merger will enable Cray to increase the availability of high-performance parallel applications and renew industry interest in advancing the capability of supercomputer applications. SGI's excellent workstation and server market coverage will free Cray to focus on the high-end of supercomputing. The price/performance of the highest-end supercomputing systems will equal that of high-end desktop systems, but with capabilities that are orders of magnitude greater.

This paper will define Cray's software goals and strategies for continued supercomputing improvements. The paper will give the status of the current Cray operating systems, programming environments, and I/O features, and describe plans for long-term support and innovative enhancement of these current products.

2 Goals and Strategies

2.1 Long-Term Commitments

Cray is committed to maintaining and improving application and systems-code portability while providing industry-leading scalable performance. The SGI merger significantly augmented Cray's long-term emphasis on open systems software environments.

Cray's existing systems follow X/Open, POSIX, Fortran 90, networking, OSF, and other standards. This provides an excellent base for establishing compatibility among Cray and SGI platforms. This existing compatible base also includes key multiple-vendor-platform products which Cray will deploy on SGI platforms and vice versa. (Examples include NQE, CF90, LibSci, and OpenVault.) Cray systems and programming environment software has already been ported among PVP and MPP systems, which have greater differences than those between Cray MPP and Scalable Node systems. In short, Cray is well prepared to provide common features among its PVP, MPP, and Scalable Node systems.

Development of and support for the UNICOS and UNICOS/mk operating systems is ongoing. UNICOS platforms are stable and feature-rich. Cray will continue to make improvements in the recoverability and resiliency of UNICOS, while providing new hardware support and adding key IRIX features. UNICOS/mk is very young, with 18 systems shipped as of CUG, and many releases planned to augment it in order to achieve UNICOS equivalence. UNICOS/mk development is proceeding according to the plan; Cray will provide the full feature set in the plan. Cray is currently concentrating on UNICOS/mk stability.

Cray is making plans to provide its tested and proven "data center quality" software environment and premiere service across all its platforms: PVP, MPP, and Scalable Node.

2.2 Integrated Systems and Software

Cray will add UNICOS reliability and scalability features to IRIX, while continuing to develop UNICOS enhancements and adding key IRIX features to UNICOS (for example, OpenVault, BDS, and XFS). The result will be a common set of standard and enhanced features that will form a common Application Programmer Interface (API) among the PVP, MPP, and Scalable Node systems.

To integrate the hardware, SGI-Cray will provide common peripherals that will allow customers to reuse key GigaRing tape and disk peripherals on larger Scalable Node systems. SGI-Cray will provide excellent file and data sharing capabilities with protocols such as Bulk Data Service (BDS), DCE/DFS, and NFS.

Key Cray and SGI enhancements will be common across the platforms. Examples include CF90, NQE, distributed DMF, `libio` (FFIO), UNICOS batch and tape features, common C++ front ends, a common LibSci subset, Message Passing Toolkit (MPT), OpenVault (multihost tape drive management), and XFS (SGI's file system).

2.3 Cross-Platform Product Development Techniques

Cray and SGI will apply appropriate new work to all platforms: UNICOS, UNICOS/mk, and IRIX. Cray will continue to prepare a software base for common code while proceeding to incorporate portable technologies. An example is the conversion of DMF to use the DMIG file-system interface standard so that DMF will be portable across DMIG-compatible file systems, such as XFS and possibly Veritas. Cray is using common compiler front ends, libraries, and tools, and will avoid processes that deliver to a single platform.

Cray will repeat the techniques and style used in the UNICOS to UNICOS/mk migration. The result will be increased portability and flexibility, and a high-degree of integration across the platforms.

2.4 The Cellular IRIX Decision

Should Cray and SGI use UNICOS/mk or Cellular IRIX on SN1 and SN2?

> SN0 is the code name for the SGI-Cray Origin2000 product released in October 1996. SN1 and SN2 are the code names for the successor products, which will augment Origin2000 and eventually replace the Cray PVP and MPP product lines.

Cray and SGI have two scalable software technologies that could run on SN1 and SN2: UNICOS/mk and Cellular IRIX.

Both technologies scale well. IRIX has a MIPS Absolute Binary Interface (ABI) that is compatible with Origin2000 and POWER CHALLENGE. It has advanced graphics and virtual memory features. There is a clear path for adding the UNICOS programming interface (API) to Cellular IRIX.

If Cray and SGI adopted UNICOS/mk for SN1, maintenance of the required MIPS ABI, advanced graphics features, virtual memory, and other SGI features would be at risk. The better choice appears to be to deploy Cray developers to focus on enhancing scalability in Cellular IRIX rather than the larger task of developing a MIPS ABI in UNICOS/mk.

The decision is therefore clear: Cellular IRIX is the superior choice for the SN1 and SN2 operating system.

3 Software Status

3.1 UNICOS Status

Overall UNICOS system availability is 99.59% for 827 systems in the field.

Cray produced a dramatic improvement in software MTTI on all of its UNICOS platforms. The mean time between software interrupts is now greater than a year on most platforms. System MTTI (including hardware interrupts) is lower, but also improving.

UNICOS 8.0 and UNICOS 9.0 have similar rates of weighted Software Problem Reports (SPRs) per system-month of operation, when measured in weeks since release. Fifteen percent of the systems in the field are running UNICOS 9.0 (typically the larger and newer systems); 85% are running UNICOS 8.0 (typically the smaller CRAY EL and CRAY J90 systems); and 15% are running UNICOS 7.0 (typically the older CRAY Y-MP and CRAY X-MP systems). Cray expects that most CRAY J90, CRAY C90, and CRAY T90 systems will run UNICOS 9.0 or UNICOS 10.0 by the end of 1997, as support for older releases ends.

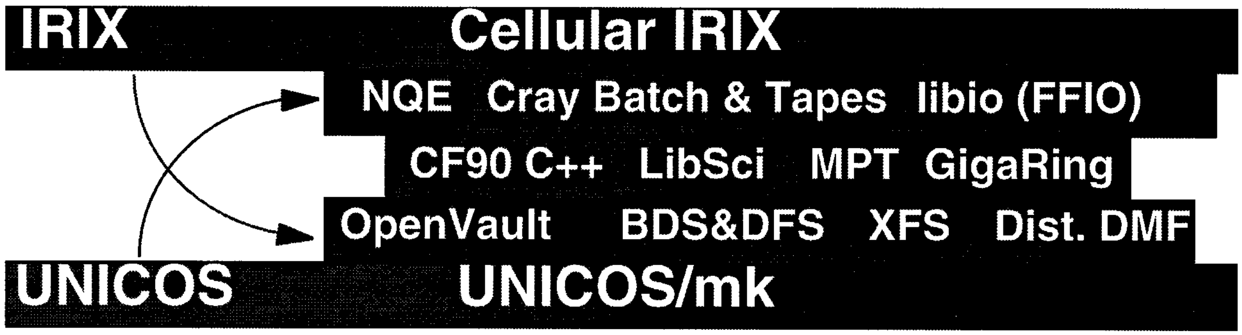

Figure 1: Common UNICOS and IRIX Features

Table 1. Software MTTI Hours

Platform	Software MTTI	System MTTI
CRAY T90	11,000 hours	600 hours
CRAY C90	22,000 hours	800 hours
CRAY J90	17,000 hours	2,300 hours
CRAY T3D	8,000 hours	1,200 hours

(13 week sample as of September 1, 1996)

3.2 UNICOS/mk Status

Cray shipped over 18 production CRAY T3E systems (10 to 256 PEs for a total of over 1,500 PEs) before the fall CUG. The first production shipment was in August 1996. The performance results have been excellent—typically 3 to 5 times faster than the CRAY T3D system, with good scaling.

System stability is improving with limited initial I/O configurations. UNICOS/mk stress testing is underway, and Cray has seen a high rate of progress on stability over the last few months. Few problems have been reported on CRAY T3E compilers and libraries.

3.3 CRAY T90-IEEE Performance

Programming Environment (PE) 2.0 was the first CF90 IEEE release. It provided equal or faster performance than that of CF77 Cray Floating Point compiled codes on Cray's large set of test suites. Some key applications, however, are not included in these test suites.

PE 3.0 will seek to improve IEEE performance with new inlining and scalar optimizations.

3.4 CRAY J90se Performance

The CRAY J90se CPUs are running with 200 MHz scalar and100 MHz vector units with the cache enabled. The performance is as expected. For example, the CRAY J90s showed a 22% improvement over the CRAY J90 performance on Livermore Loops harmonic mean. The peak (vector-dominated) speed is the same, as expected.

3.5 CRAY T3E Performance

CRAY T3E performance is typically 3 to 5 times higher than CRAY T3D performance. A streams benchmark shows 500-600 MByte/s of local memory bandwidth and 450 MByte/s of stride-independent bandwidth (gathers using E-registers). The scalability of the programs on CRAY T3E systems exceeds the already excellent scalability of programs on CRAY T3D systems.

3.6 Current Development Status

Almost all Cray developers are working on CRAY J90s, CRAY T90, and CRAY T3E projects, with a few developers working on SN projects. Developers are preparing to apply new work to all platforms. Cooperative relationships among Cray and SGI developers are progressing well, with an excellent spirit of choosing the best features and implementations from the common pool of Cray and SGI technology.

4 Software Plans

4.1 Programming Environments Group

Cray will provide common key languages, features, and libraries across its PVP, MPP, and MIPS product lines. Cray will offer CF90 and common C++/C front ends on all platforms (PVP, MPP, MIPS). This will include common programming models (AutoTasking, message-passing, SHmem, etc.) and common key libraries (Fortran I/O (FFIO), `libf`, `libsci`, etc.).

4.1.1 PVP and MPP Programming Environments

Cray released PE 2.0 (CF90 and C++/C) in 1996 on Cray PVP, CRAY T3D, and CRAY T3E platforms. On the CRAY T3D platform, this was a 4Q96 release that included double the

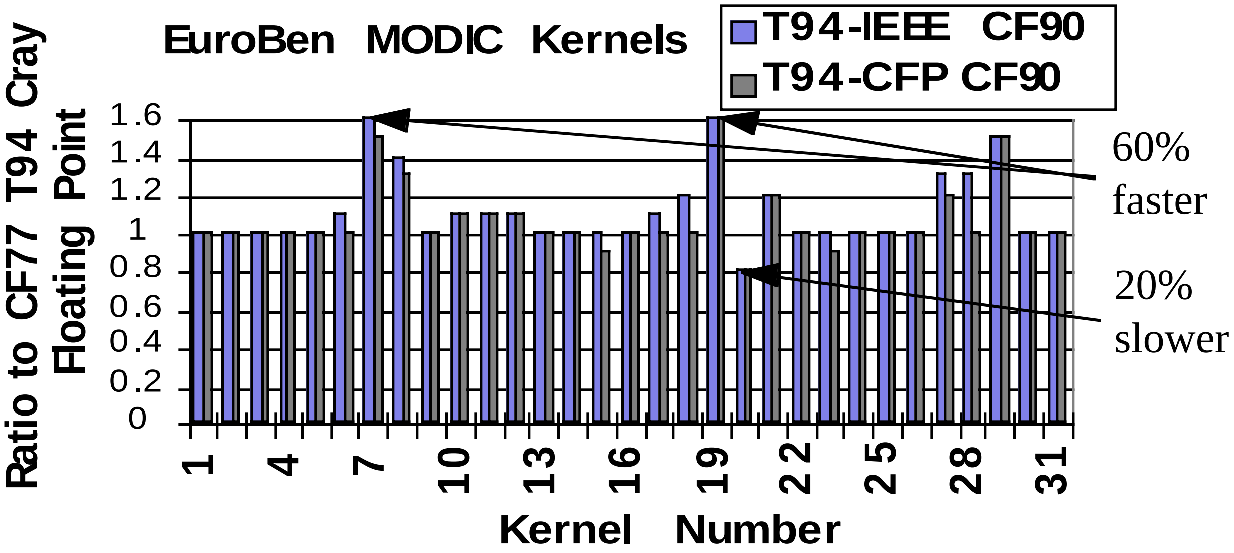

Figure 2: CRAY T94 IEEE Benchmarks

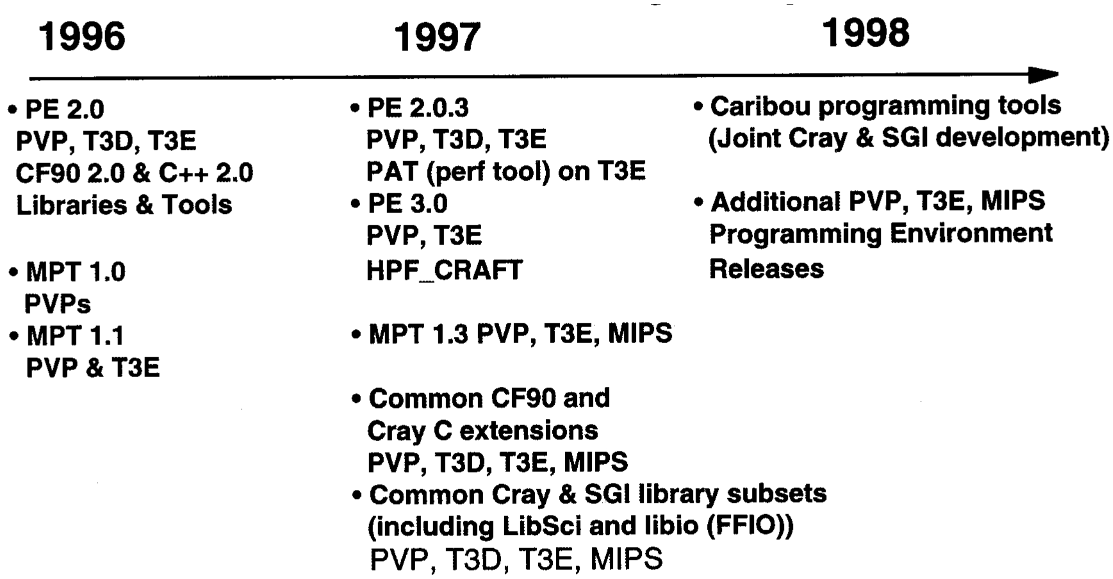

Figure 3: Programming Environment Timeline

performance for square root and other `libm` improvements. It also included global I/O and the full CF90 language, which provides CRAY T3D and CRAY T3E consistency. PE revision 2.0.3 will be released in 1Q97 on Cray PVP, CRAY T3D, and CRAY T3E platforms. PE 2.0 also supports systems with mixed CRAY T90-IEEE/Cray floating-point CPUs.

Cray plans to release PAT (performance analysis tool) for the CRAY T3E system in 1997. PAT will augment the analysis available in Cray's Apprentice tool and will be a forerunner for the Caribou tools that are planned for the Scalable Nodes.

PE release 3.0 is planned for 2Q97 on Cray PVP and CRAY T3E platforms. It will include CF90 scalar and inlining optimizations, along with C++ performance enhancements and standards improvements. PE 3.0 will support for HPF_CRAFT on CRAY T3E systems. (HPF_CRAFT may require a separate license.)

4.1.2 MIPS Programming Environments

The initial Cray-compatibility MIPS programming environment is targeted for the MIPS F90 alpha test in 3Q97. It integrates Cray CF90 front-end technology with MIPS optimization and code generation. It will include Cray Fortran I/O (including FFIO) and the LibSci subset. This LibSci subset is a joint Cray-SGI project with combined scientific libraries that are scalable and high in performance. This subset will include many scalable routines from the CRAY T3E platform.

This release will also include the initial version of Caribou—a jointly developed set of new programming tools.

4.2 Operating Systems Group

Cray will provide long-term UNICOS and UNICOS/mk development and support, with releases for both platforms planned throughout this decade.

4.2.1 UNICOS Releases 1996 to 2000+

- UNICOS 8.0.4.4 July 1996
  Final 8.0 update
- UNICOS 9.0 (major release) 1995-1996
  EL, J90, C90, T90
- UNICOS 9.1 (T90-IEEE): March 1996
- UNICOS 9.2 & 9.3
  J90se-GR and T90-GR 1Q97
- UNICOS 10.0
  Planned for 1997 release
  C90, T90, T90-IEEE, T90-GR, J90, J90se
  Will also support J90++ and PV+
  Initial IRIX features (OpenVault, BDS, ...)

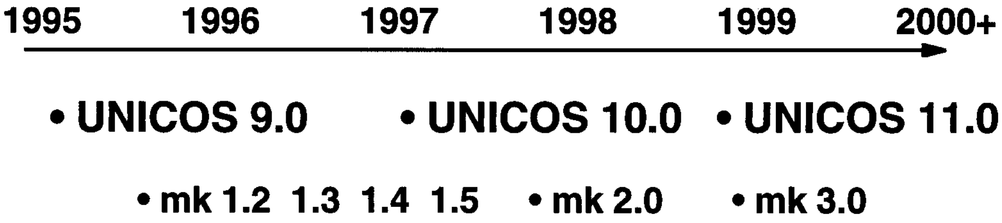

Figure 4: UNICOS Timeline

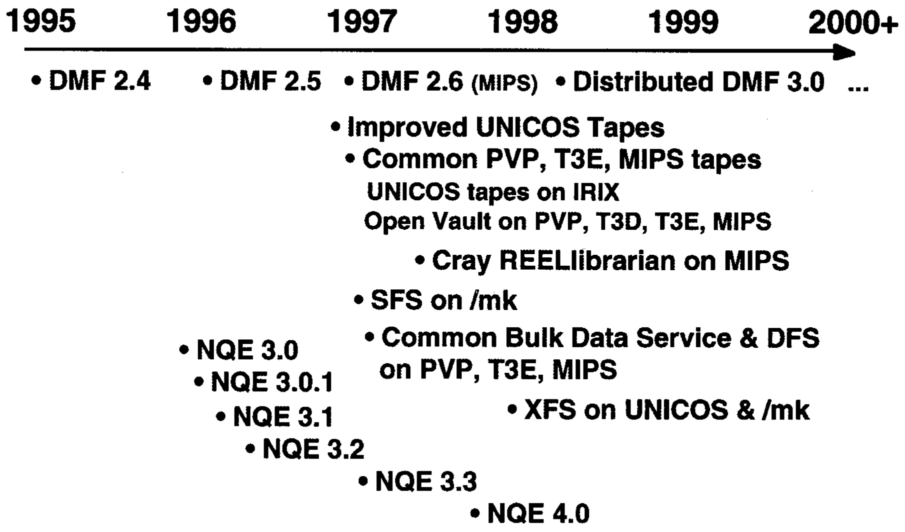

Figure 5: IO Group Timeline

- UNICOS 11.0 planned for 1999-2000
  Additional IRIX features (XFS, ...)
  Continuing reliability and resiliency improvements

4.2.2 UNICOS/mk Releases

- UNICOS/mk 1.2: 3Q96
  Initial multiple PE OS
- UNICOS/mk 1.3: 4Q96
  NQE 3, initial tape support, process migration, configuration support
- UNICOS/mk 1.4: 1Q97
  Timesharing and swapping parallel applications, political scheduling, DMF, full GigaRing support, disk quotas
- UNICOS/mk 1.5: 2Q97
  Checkpoint-restart, URM, DCE/DFS, complete resource limits, CRL
- UNICOS/mk 2.0: 1998
- UNICOS/mk 3.0: 1999+

4.3 I/O Group

The IO group will release many existing UNICOS products in common across the PVP, MPP, and MIPS product lines, including DMF, NQE, and UNICOS tape and batch features. This group will also add key IRIX features, such as OpenVault and the XFS file system, to UNICOS.

Key IO releases:

- DMF 2.6—2Q97 on MIPS
- DMF 3.0—1998 on PVP, T3E, and MIPS
  Common distributed architecture
- NQE 3.2—4Q96 (3.3 and 4.0 in 1997)
  Improved T90 recovery, T3D rolling, IRIX checkpointing, Origin 2000
- Tapes—2Q97 on UNICOS, UNICOS/mk, and IRIX
  OpenVault (multihost tape drive management)
- Bulk Data Services (BDS) & DFS
  Fast multihost (PVP, T3E, MIPS) file systems
- XFS (IRIX file system) on UNICOS and UNICOS/mk

5 Summary

Cray and SGI will provide long-term UNICOS and UNICOS/mk support. These operating systems will be competitive throughout this decade. Cray will boost the CRAY T90 speed, ensuring CRAY T90 and PV+ systems remain the platforms of choice for many high-end vector-SMP codes throughout this decade. Several speed improvements are also in the pipeline for CRAY J90 systems, including J90se and J90++. These systems will offer the best price/performance for vector codes. The CRAY T3E systems are the most scalable machines available, with excellent MPP performance and price/performance that scales to teraflops of computing speed, terabytes of DRAM, and gigabytes per second of I/O.

Cray will maintain its focus on supercomputing, continuing to improve performance and price/performance. Cray and SGI are implementing software plans for merging UNICOS and IRIX features, which will protect existing customer investments in Cray technology.

This strategy will allow Cray to invest in high-performance technologies, where Cray differentiates its products, while using the high-volume base of SGI technology and applications. The result will be reduction in costs for developing high-end supercomputing hardware and software, along with a broadening of the binary- and source-compatible application base.

CRAY
RESEARCH
A Silicon Graphics Company